LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day

Year: 2023 | Citations: 1223 | Authors: Chunyuan Li, Cliff Wong, Sheng Zhang

Abstract

Conversational generative AI has demonstrated remarkable promise for empowering biomedical practitioners, but current investigations focus on unimodal text. Multimodal conversational AI has seen rapid progress by leveraging billions of image-text pairs from the public web, but such general-domain vision-language models still lack sophistication in understanding and conversing about biomedical images. In this paper, we propose a cost-efficient approach for training a vision-language conversational assistant that can answer open-ended research questions of biomedical images. The key idea is to leverage a large-scale, broad-coverage biomedical figure-caption dataset extracted from PubMed Central, use GPT-4 to self-instruct open-ended instruction-following data from the captions, and then fine-tune a large general-domain vision-language model using a novel curriculum learning method. Specifically, the model first learns to align biomedical vocabulary using the figure-caption pairs as is, then learns to master open-ended conversational semantics using GPT-4 generated instruction-following data, broadly mimicking how a layperson gradually acquires biomedical knowledge. This enables us to train a Large Language and Vision Assistant for BioMedicine (LLaVA-Med) in less than 15 hours (with eight A100s). LLaVA-Med exhibits excellent multimodal conversational capability and can follow open-ended instruction to assist with inquiries about a biomedical image. On three standard biomedical visual question answering datasets, LLaVA-Med outperforms previous supervised state-of-the-art on certain metrics. To facilitate biomedical multimodal research, we will release our instruction-following data and the LLaVA-Med model.